Plataforma de Aprendizado de Máquina para Detecção e Monitoramento de Alunos com Risco de Evasão

Carlos Antonio R. Beltrán [1], João C. Xavier-Júnior[1], Cephas A. da S. Barreto[2], Carlos A. de O. Neto[2]

[1]Instituto Metrópole Digital - Universidade Federal do Rio Grande do Norte
Natal, Rio Grande do Norte – Brasil

[2]Departamento de Informática e Matemática Aplicada -
Universidade Federal do Rio Grande do Norte
Natal, Rio Grande do Norte – Brasil

carvan1521@gmail.com, jcxavier@imd.ufrn.br,

cephasax@gmail.com, carlosantoniooln@gmail.com

Abstract. *Machine Learning Techniques (AM) have been applied to a wide range of problems in the real world (e.g., classification, regression, clustering, etc). Based on this fact, this work proposes the development of a machine learning tool for detecting and monitoring higher education students at risk of dropout. As dropout in higher education is a global problem characterized by interruption of school registration for a period of time, an academic database owned by the Los Angeles Catholic University of Chimbote (ULADECH) located in Peru was used to validate our proposal.*

Resumo. *Técnicas de Aprendizado de Máquina (AM - do inglês - Machine Learning) têm sido aplicadas aos mais variados problemas do mundo real, principalmente devido ao grande potencial e diversidade das mesmas. Baseado nessa realidade, este trabalho propõe o desenvolvimento de uma plataforma de Aprendizado de Máquina para detecção e monitoramento de alunos do ensino superior em risco de evasão. Como a evasão no ensino superior é um problema global caracterizado pelo abandono de curso, foram utilizados os dados de alunos da Universidade Católica Los Angeles de Chimbote (ULADECH) localizada no Peru para validar a nosso proposta.*

1. Introdução

O termo evasão pode apresentar conceitos bastante diversos. Essa constatação se baseia no número de definições que pode ser encontrado na literatura [Figueiredo and Salles 2017]. Para este trabalho, foi adotada a definição de Johann (p. 65) [Johann 2012], que descreve a evasão como sendo "um fenômeno caracterizado pelo abandono do curso, rompendo com o vínculo jurídico estabelecido, não renovando o compromisso ou sua manifestação de continuar no estabelecimento de ensino".

Segundo Baggi [Baggi and Lopes 2011], esse é um problema que preocupa as instituições de ensino públicas e particulares, pois a saída de alunos provoca graves consequências sociais, acadêmicas e econômicas. Ainda segundo os autores, a evasão média ocorrida nas Instituições de Ensino Superior (IESs) do Brasil, entre os anos de 2000 e

DOI: 10.5753/cbie.sbie.2019.1591

2005, foi de 22%, e atingiu 12% nas públicas e 26% nas particulares. Além disso, tal pesquisa ainda revelou que são poucas as instituições que possuem um programa institucional regular de combate à evasão, com planejamento de ações, acompanhamento de resultados e coleta de experiências bem sucedidas.

Além do Brasil, também outros países vizinhos da América Latina sofrem com a evasão em suas IESs. No Peru, por exemplo, há um distanciamento grande entre o conteúdo ensinado nas escolas e o que é exigido como fundamento básico nas universidades. Essa defasagem de conteúdo contribui para o alto número de evasões dentro das IESs no Peru [Tudela 2014]. Segundo fontes, entre 40 e 50 mil jovens abandonam seus estudos anualmente. Desse grupo de jovens, 70% corresponde a estudantes de instituições particulares e 30% de públicas [Plasencia 2011].

Diante de tal cenário, é preciso investigar os possíveis fatores que levam os alunos ingressantes no ensino superior a se evadirem. Motivados por esse desafio, este trabalho propõe o desenvolvimento de uma plataforma de Aprendizado de Máquina que possibilite a detecção e o monitoramento de alunos com risco de evasão. O protótipo da plataforma em questão utiliza os dados acadêmicos e socio-econômicos dos alunos de uma universidade localizada no Peru. Essa universidade é uma das muitas universidades peruanas que tem enfrentado o desafio de descobrir a origem e os motivos causadores do elevado número de evasão.

De fato, a evasão em IESs já vem sendo estudada por alguns pesquisadores [Lanes and Alcântara 2018], [Digiampietri et al. 2016], [Costa Gonçalves et al. 2018], [Manhães et al. 2011] e [Paz and Cazella 2017]. Contudo, diferentemente dos demais trabalhos, este apresenta uma plataforma de Aprendizado de Máquina capaz de pré-processar os dados socio-econômicos e acadêmicos dos alunos, e, a partir deles, prever os alunos em risco de evasão (através de modelos de AM). Além disso, ela oferece recursos administrativos capazes de monitorar tais alunos em risco, indicando tutores para acompanhá-los, assim como material extra para estudo através de gamificação.

2. Conceitos Relacionados

Como mencionado anteriormente, as técnicas de Aprendizado de Máquina têm sido aplicadas aos mais variados problemas do mundo real na última década [Faceli et al. 2011, Witten et al. 2016]. Contudo, é importante conhecer bem os dados disponíveis (base de dados), para depois aplicar corretamente as técnicas de AM.

Em geral, essas técnicas podem ser dividas em dois grupos: as preditivas (e.g., classificação e regressão) e as descritivas (e.g., agrupamento e associação). O objetivo das técnicas preditivas é encontrar uma função (hipótese), que a partir dos dados de treinamento possa ser utilizada para prever um rótulo ou valor que caracterize um novo modelo. Já o objetivo das técnicas descritivas é explorar ou descrever um conjunto de dados, não utilizando para tal o rótulo ou atributo de saída. Tais técnicas seguem o paradigma de aprendizado não supervisionado (sem a presença de um supervisor).

É importante ressaltar que as técnicas de Aprendizado de Máquina Supervisionado e não Supervisionado são muitas. Por isso, os esforços de pesquisa foram concentrados em técnicas de Classificação (Supervisionadas) e Agrupamento ou *Clustering* (não Supervisionadas). Dessa forma, serão descritas todas as técnicas utilizadas nos experimentos nas subseções seguintes.

2.1. Técnicas de Classificação

Essas técnicas são aplicadas em problemas onde as instâncias de uma base de dados, representados por um conjunto de atributos, precisam ser enquadrados em um conjunto pré-definido de possíveis rótulos (classes). Há um bom número de algoritmos de classificação propostos na literatura [Faceli et al. 2011, Witten et al. 2016]. As seis técnicas utilizadas neste trabalho são descritas a seguir:

- AdaBoost: implementação do algoritmo Boosting [Schapire 1999];
- Bagging: algoritmo que utiliza múltiplas versões dos dados para treinar classificadores de mesmo tipo [Breiman 1996].
- IBk: implementação do algoritmo k-Nearest Neighbor (k-NN);
- J48: implementação do algoritmo Árvore de Decisão;
- Naive Bayes (NB): algoritmo estatístico baseado no teorema de Bayes;
- MultiLayer Perceptron (MLP): tipo de Rede Neural que é baseada no modelo biológico do sistema nervoso;

Para analisar o desempenho de cada uma das técnicas de classificação foi utilizada uma medida de desempenho: F-measure. Ela é mais indicada para bases desbalanceadas, e pode ser definida pela seguinte equação:

$$F - measure = \frac{(2 * precision * recall)}{(precision + recall)} \quad (1)$$

2.2. Técnicas de Agrupamento

Essas técnicas são aplicadas em problemas onde a base de dados não possui rótulos (classes). Dessa forma, é necessário aplicar técnicas baseadas em distância, densidade ou outra heurística, que possam descrever o padrão dos dados. Há uma grande variedade de algoritmos de agrupamento que podem ser encontrados na literatura [Faceli et al. 2011]. Neste trabalho foram utilizados os seguintes algoritmos: k-Means, Expectation Maximization (EM) e Hierárquico Aglomerativo. Para validar as partições geradas pelos três algoritmos de *clustering* foram utilizadas duas medidas de validação muito conhecidas nessa área de validação, que são: Davies-Bouldin (DB) e Silhueta [Halkidi et al. 2002].

3. Trabalhos Relacionados

Em geral, a aplicação de técnicas de Aprendizado de Máquina como ferramenta útil no combate a evasão escolar tem recebido notoriedade nos últimos anos. Contudo, por serem muitas técnicas, como já discutido, os trabalhos podem aplicá-las tanto na predição quanto na descrição de padrões de evasão. Como este trabalho tem a natureza de detecção (classificação) de alunos com risco de se evadir, serão considerados somente em trabalhos que fazem uso das mesmas técnicas de AM.

No trabalho de [Lanes and Alcântara 2018], os autores utilizaram dados acadêmicos referentes ao primeiro ano dos cursos de graduação para identificar alunos em risco de evasão. Os dados de 12 cursos de graduação foram extraídos do sistema acadêmico da Universidade Federal do Rio Grande (FURG) no período de 2012 até 2017. O algoritmo J48 (Árvore de Decisão) foi utilizado como gerador de regras para o sistema de classificação. O mesmo método e objetivos também podem ser vistos no trabalho de [Paz and Cazella 2017].

Em um outro trabalho [Digiampietri et al. 2016], os autores também utilizaram somente os dados referente ao desempenho nas disciplinas do primeiro ano de curso. Foram analisados os dados extraídos dos históricos de mais de 1000 alunos do Bacharelado de Sistemas de Informação da EACH-USP entre os períodos 2005 e 2015. Para este trabalho, foi utilizado o classificador Rotation Forest, que por padrão aplica seleção de atributos a partir do *Principal Component Analysis* (PCA).

Já no trabalho de [Costa Gonçalves et al. 2018] foram utilizados três algoritmos (Naive Bayes, Support Vector Machine e J48) para identificar quais alunos da graduação no Instituto Federal do Maranhão (IFMA) eram os mais propensos a abandonar os estudos. Além disso, três abordagens de seleção de atributos foram testadas (Manual, Correlação e Ganho de Informação). Utilizou-se um dataset composto de 40 atributos e 574 instâncias para testar os classificadores.

Por último, Manhaes [Manhães et al. 2011] apresentou um estudo utilizando dados acadêmicos de alunos de graduação de Engenharia da Escola Técnica da Universidade Federal do Rio de Janeiro (UFRJ) para identificar precocemente aqueles com risco de evasão. A base de dados possuia 543 alunos que concluíram o curso e 344 que se evadiram. Nesse trabalho foram utilizados 10 algoritmos (OneR, JRip, DecisionTable, SimpleCart, J48, RandomForest, SimpleLogistic, MultilayerPerceptron, NaiveBayes e BayesNet). Os resultados mostraram que utilizando as primeiras notas semestrais dos calouros é possível identificar com precisão de 80% a situação final do aluno no curso.

Diferentemente dos trabalhos aqui discutidos, este trabalho tem o objetivo de propor uma plataforma de Aprendizado de Máquina capaz de recuperar dados acadêmicos e socio-econômicos de alunos dos mais variados cursos de graduação, processar esses dados, aplicar um modelo de AM já treinado, identificar os alunos com risco de evasão e monitorá-los através de procedimentos pedagógicos.

Dessa forma, fica claro que não se trata apenas de um teste com diferentes algoritmos de classificação (classificador base e *Ensembles*) para selecionar o modelo mais indicado a ser utilizado sobre os dados. Embora a análise dos classificadores, feita também pelos trabalhos anteriores, seja importante, há outros aspectos que precisam ser observados, principalmente aqueles relacionados ao perfil (acadêmico e social) dos alunos em risco de evasão.

4. Metodologia

A metodologia para construção do modelo de Aprendizado de Máquina a ser utilizado na plataforma proposta segue uma lógica sequencial, onde 4 passos são realizados com o objetivo de encontrar o modelo que será responsável por informar se um aluno possui ou não risco de evasão. A Figura 1 demonstra, de forma simplificada, os passos utilizados para construção do modelo citado. Desta forma, é possível entender os passos usados na metodologia como:

- **passo 1:** aplicação da correlação de Pearson [Witten et al. 2016] sobre a base original (v1, com 36 atributos) para seleção dos melhores atributos (excluídos atributos com correlação maior que 70%). Como resultado foi obtida uma segunda versão da base com 19 atributos, denominada v2;
- **passo 2:** criação de diferentes partições através da utilização de algoritmos de agrupamento (EM, HA e k-Means). De acordo com as métricas de validação

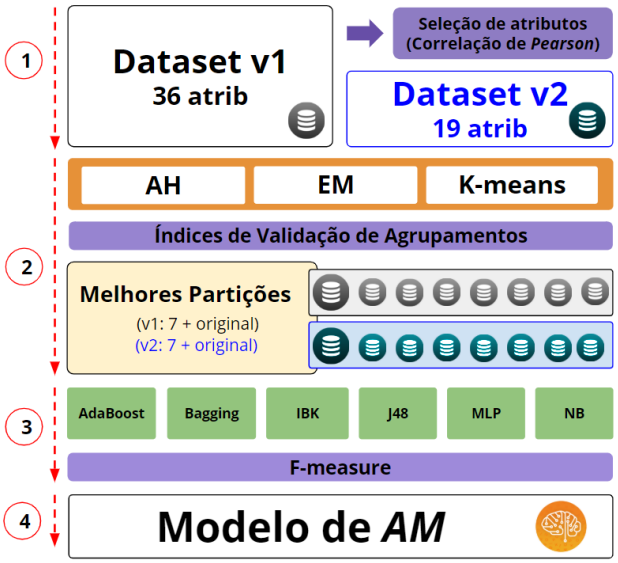

Figura 1. Metodologia para Construção do Modelo de AM

Silhouette e DB, foram escolhidas as 7 melhores partições de cada base (v1 e v2), perfazendo um total de 14 (quatorze). Essas partições se juntaram às bases v1 e v2 e foram utilizadas como entrada do passo seguinte. A utilização de algoritmos não supervisionados foi necessário para gerar um número maior de bases, e assim poder analisar estatisticamente o desempenho dos classificadores analisados.

- **passo 3:** aplicação de seis métodos de classificação sobre as 16 bases de entrada, v1 e suas 7 melhores partições e v2 e suas 7 melhores partições. Os métodos AdaBoost, Bagging, IBK, J48, MLP e Naive Bayes foram utilizados e, avaliados segundo a métrica F-measure.
- **passo 4:** escolha do melhor modelo de acordo com os testes estatísticos (Friedman e Post-hoc) realizados sobre os resultados de acurácia de cada método.

Após escolhido o modelo, este foi então treinado com 2.696 instâncias. Em seguida, foi testado com todos os alunos que estão cursando no momento (3.393), apresentando resultados maiores que 90%. Isto significa que, de acordo com o modelo, há um alto risco de evasão para mais de 90% dos alunos que estão cursando suas graduações. Este resultado possui implicações que serão discutidas nas seções 6.3 e 7.

O modelo em questão foi então incorporado à plataforma educacional proposta, e possibilitará identificar grupos de alunos com alta probabilidade de evasão (antes que ela ocorra), o que pode ajudar significativamente nas decisões de suporte educacional e social tomadas junto a esses alunos.

5. Resultados Experimentais

A base de dados utilizada neste trabalho foi extraída do sistema acadêmico de uma universidade católica localizada no Peru, e conta com dados socio-econômicos e acadêmicos de alunos de graduação matriculados entre os períodos acadêmicos de 2010 e 2017. Após a etapa de pré-processamento, a referida base ficou com 36 atributos e 6.089 instâncias, das quais 873 são de alunos que concluíram, 1.823 são de alunos que se evadiram, o restante (3.393) são de alunos que estão cursando. Esta base será chamada de v1.

Tabela 1. Descrição dos Atributos.

#	Atributo	Tipo Atributo	#	Atributo	Tipo Atributo
1	**sexo**	binário	19	rCatolico	binário
2	**idade**	numerico	20	rOutra	binário
3	zmResidencial	binário	21	rEvangelico	binário
4	zmAssentamento	binário	22	rAgnostico	binário
5	zmUrbanicao	binário	23	cursograduacao	nominal
6	zmCascourbano	binário	24	**modalidade**	binário
7	zmPovoJoven	binário	25	**periodoIngreso**	numérico
8	zmOtros	binário	26	**notaExameAdmissao**	numérico
9	zmZonaRural	binário	27	**totalNiveisCurso**	numérico
10	trabalha	nominal	28	**totalCreditosCurso**	numérico
11	**ecSolteiro**	binário	29	**totalNiveisCursados**	numérico
12	**ecCasado**	binário	30	**porcentagemNC**	numérico
13	ecDivorciado	binário	31	**totalCreditosCursados**	numérico
14	**ecDesconhecido**	binário	32	**porcentagemCC**	numérico
15	ecSeparado	binário	33	**periodoUltimaMatricula**	numérico
16	ecUestavel	binário	34	**ultimoNivelEstudado**	numérico
17	ecViuvo	binário	35	**promedioUltimoPeriodo**	numérico
18	**escola**	nominal	36	**class**	nominal

A Tabela 1 apresenta todos os 36 atributos e seus respectivos tipos. Note que os tipos binários representam todos os possíveis valores nominais para um determinado atributo (e.g. religiaoCatolica - 1 representa sim e 0 representa não). Além desta base, também foi utilizada uma versão menor, chamada de V2, com 19 atributos (em negrito na referida tabela) selecionados através da Correlação de Pearson, e outras 14 versões geradas a partir de algoritmos de *Clustering*, como descrito na seção 4.

5.1. Análise das Técnicas de AM

A Tabela 2 apresenta os valores médios da métrica F-measure para seis diferentes classificadores (colunas de 3 a 8). Note, que o valor médio para cada base de dados é composto pela média de três metodologias de treinamento/teste (10 fold cross validation, 70/30 e 60/40, e.i., onde a primeira parte representa o percentual da base utilizado para treinamento, e a segunda o percentual para teste). O uso de diferentes metodologias deixa mais robusta qualquer análise feita sobre os classificadores.

Com relação aos classificadores, é importante ressaltar que foram feitos vários experimentos, sendo reportados somente os melhores resultados. O IBK com 5 vizinhos, o MLP-t com 39 neurônios na camada escondida, o AdaBoost (ADA) e o Bagging (BAG), ambos, utilizando o MLP-a com 19 neurônios na camada escondida como classificador base foram as melhores configurações para tais algoritmos. Para os demais, Naive Bayes (NB) e Árvore de Decisão (J48), foram utilizadas as configurações padrões do WEKA para cada um deles.

Ainda com relação a Tabela 2, é possível visualizar que o BAG (MLP-a) obteve os melhores valores médios de F-measure, ficando J48 na segunda posição e MLP-t na terceira. Também é possível perceber que NB obteve os piores resultados dentre os seis métodos de classificação (Ensembles e Classificadores base).

Contudo, visando apresentar uma análise estatística mais robusta, foi aplicado o

Tabela 2. Média dos valores de F-measure para os Classificadores.

Id	Bases	IBK(5)	J48	MLP(t)	NB	ADA(MLP-a)	BAG(MLP-a)
1	V1_Original	0,7100	0,8263	0,7810	0,7360	0,7923	0,8030
2	V1_EM-2k	0,9977	0,9977	0,9980	0,9940	0,9980	0,9980
3	V1_EM-3k	0,9800	0,9917	0,9907	0,9600	0,9903	0,9930
4	V1_HA-2k_cLink	0,9930	0,9903	0,9920	0,9633	0,9923	0,9927
5	V1_HA-3k_cLink	0,9800	0,9713	0,9830	0,8330	0,9770	0,9847
6	V1_HA-4k_cLink	0,9787	0,9703	0,9780	0,8240	0,9773	0,9833
7	V1_kM-2k_37seed	0,9997	1,0000	1,0000	1,0000	1,0000	1,0000
8	V1_kM-3k_37seed	0,9680	0,9823	0,9930	0,9763	0,9933	0,9980
9	V2_Original	0,7627	0,8263	0,7997	0,7400	0,8037	0,8110
10	V2_EM-2k	0,9870	1,0000	0,9997	0,9707	0,9997	1,0000
11	V2_EM-3k	0,9703	0,9943	0,9923	0,9533	0,9940	0,9923
12	V2_HA-2k_cLink	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
13	V2_HA-3k_cLink	0,9837	0,9680	0,9833	0,9387	0,9770	0,9843
14	V2_HA-4k_cLink	0,9837	0,9687	0,9840	0,9387	0,9763	0,9850
15	V2_kM-2k_37seed	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
16	V2_kM-3k_37seed	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Média		0,9559	0,9680	0,9672	0,9268	0,9670	**0,9703**
Desv. Padrão		0,0869	0,0566	0,0695	0,0911	0,0666	0,0641

Teste de Friedman para verificar se há diferença estatística entre os resultados dos classificadores. O referido teste gerou um resultado de $p - value = 0{,}0003$, configurando-se, dessa forma, diferença estatística entre os métodos. Logo, foi utilizado um teste post-hoc (Nemenyi) para analisar os métodos dois a dois. O teste Nemenyi apontou um $p-value = 0{,}0350$ entre BAG (MLP-a) e IBK (5) e $p - value = 0{,}0002$ entre BAG (MLP-a) e NB, o que caracteriza diferença estatística entre esses métodos.

Porém, não houve diferença estatística entre BAG (MLP-a) e os demais métodos (ADA, MLP e J48). Dessa forma, de acordo com o teste Nemenyi, como não houve diferença estatística entre os métodos, qualquer um poderia ser utilizado como modelo. Inicialmente, será utilizado o próprio Bagging, mas outros poderão ser utilizados no futuro.

6. Desenvolvimento da Plataforma Proposta

O protótipo desenvolvido caracteriza-se como um sistema educacional baseado em tecnologias para Internet, que é capaz de utilizar uma base de dados acadêmicos e socioeconômicos de alunos, e oferece funcionalidades que podem auxiliar alunos em risco de evasão. Vale salientar que a identificação de risco de evasão se dá através do modelo de AM escolhido e demonstrado nas seções anteriores. Este modelo é responsável por classificar alunos que possuem o mesmo perfil, e então apontar aqueles que estão com baixo rendimento ou com problemas (risco de se evadir).

6.1. Funcionamento da Plataforma

De forma resumida, as funcionalidades elencadas para a ferramenta são:

1. Listar todos os alunos, classificando-os em "situação normal" ou "risco de evasão";
2. Mostrar o detalhamento dos dados do aluno e confrontá-lo com os dados do perfil "situação normal" para o seu respectivo curso;

3. Mostrar os valores médios por atributo, agrupando-os (atributos) por curso e por perfil (“normal” ou “risco de evasão”);
4. Possibilitar ao gestor a realização ações diretas, tais quais:
    - Verificar detalhamento financeiro/social;
    - Verificar detalhamento de rendimento escolar;
    - Agendar reunião de orientação pedagógica;
    - Agendar conversa com o aluno;
    - Escolher e sugerir material de apoio;
    - Sugerir suporte de tutoria;
    - Notificar professores para acompanhamento aproximado;

As funcionalidades mencionadas foram desenvolvidas utilizando: o banco de dados relacional PostgreSQL (v. 10); o *framework* SpringBoot, com os módulos: JPA e WEB para configuração de acessos via API REST. Com essa estrutura é possível ter um sistema que funciona semelhante a um site, e que possui a parte de inteligência (AM) e as regras de funcionamento centralizadas em um servidor de aplicação *WEB*.

6.2. Estrutura da aplicação e Divisão de Funcionalidades

A aplicação desenvolvida foi dividida em pacotes, obedecendo à arquitetura padrão proposta pelo *framework* Spring. Desta forma, se a aplicação for pensada de forma modular, onde cada módulo é responsável por certas tarefas, pode-se ter a seguinte configuração:

- **Modelagem e acesso aos dados**: os pacotes model, repository, enums e exception são responsáveis por definir cada objeto a ser utilizado pela aplicação. Também contém as formas de acesso e recuperação dos dados no banco;
- **Regras, controle e transformação dos dados**: os pacotes converter, service e wekautils são responsáveis pelas regras de conversão e formatos de acesso a cada parte dos objetos;
- **Acesso e aprendizado de máquina**: os pacotes controller, manager e application são responsáveis por definir rotas de acesso (*endpoints*) numa API REST a ser usada pela parte visual da aplicação no navegador. Também são responsáveis por definir o fluxo de utilização do modelo para obter a situação de um aluno;
- **Recursos**: o pacote resources contém os arquivos usados pela parte da visualização (aplicação Web) e também os arquivos que dão suporte ao modelo de AM construído e exportado pelo WEKA.

Dessa forma, foi possível utilizar e comprovar a eficiência da plataforma. As funcionalidades supracitadas estão representadas através de algumas telas (*com dados fictícios*), que podem ser vistas na forma de figuras nos links:

- Tela de listagem de alunos: é possível ver todos os alunos do curso, status atual (definido pelo modelo de AM), e ações possíveis para um grupo de estudantes escolhidas em forma de botões (`bit.ly/31ZbWGN`).
- Tela de listagem de alunos por curso. Idêntica à listagem geral, mas contendo apenas alunos de um determinado curso (`bit.ly/2Z9nA4X`).
- Tela de detalhamento do aluno: é possível confrontar os índices obtidos pelo aluno e os índices médios do seu respectivo curso, a partir de diferentes perspectivas (atributos). Isto serve de subsídio para tomada de decisão mais diretas por parte das equipes pedagógica administrativa (`bit.ly/2Z6wFvm`).
- Tela de perfis médios por curso: é possível observar os valores médios por curso, e dessa forma comparar o desempenho de cada aluno (`bit.ly/2Z7LQnM`).

6.3. Conclusão Preliminar

Foi possível perceber, ao final do desenvolvimento, que há uma forte aplicabilidade da ideia geral da ferramenta em ambientes acadêmicos reais. Isso ocorre porque, para gestores, pedagogos, professores e demais responsáveis, é muito difícil prever problemas no desempenho de um aluno ou mesmo de um grupo deles. Por isso, utilizar uma plataforma que utiliza o conhecimento extraído dos próprios dados, para tomada de decisão, facilita uma das tarefas mais difíceis na educação, que é quantificar o desempenho e identificar falhas durante o processo de aprendizagem.

A partir do protótipo desenvolvido foi possível testá-lo com dados reais de alunos de graduação. Inicialmente foi detectado um percentual de aproximadamente 90% de alunos em risco de evasão. A partir dessa descoberta, através de algoritmos de *Clustering*, criou-se três grupos (baixo, médio e alto) que representam os alunos em risco de evasão. Dessa forma, os profissionais de educação poderão utilizar abordagens distintas para cada grupo de aluno em risco.

É importante ressaltar que o Peru enfrenta um grave problema de evasão nos cursos de graduação. Assim como ocorre no país, na universidade ULADECH não poderia ser diferente. Contudo, com o desenvolvimento da plataforma de AM, objetivo principal deste trabalho, várias abordagens poderão ser empregadas para diminuir esse índice tão alto. Entre elas, é possível citar as medidas que visam diminuir as retenções semestrais (período letivo), a criação de tutoria para que os alunos possam ser acompanhados por professores ou pedagogos, e por último, a criação de serviços de monitoria aplicados aos componentes curriculares que mais retem alunos em seus mais variados cursos.

7. Conclusão e Trabalhos Futuros

Este trabalho propôs o desenvolvimento de uma plataforma de Aprendizado de Máquina para detecção e monitoramento de alunos do ensino superior do Peru em risco de evasão. Para tal, foi desenvolvido um protótipo web que utiliza os próprios dados acadêmicos e socio-econômicos dos alunos para classificar os que apresentam risco de se evadir.

Visando oferecer subsídios para um bom monitoramento dos alunos em risco, várias funcionalidades foram incorporadas ao protótipo, principalmente no que diz respeito ao monitoramento baseado nos diferentes grupos de risco. Dessa forma, algumas abordagens pedagógicas puderam ser aplicadas como forma de melhorar o rendimento escolar de cada aluno, e consequentemente diminuir o alto índice de evasão.

Como trabalhos futuros, é importante avaliar quantitativamente todas as abordagens pedagógicas aplicadas, principalmente no tocante aos recursos empregados. Além disso, outras funcionalidades serão introduzidas na ferramenta, tais como: recomendação de material didático e gamificação como forma de melhoria do aprendizado.

Referências

Baggi, C. A. D. S. and Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 16:355 – 374.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Costa Gonçalves, T., Silva, J., and Andres Carmona Cortes, O. (2018). Técnicas de mineração de dados: Um estudo de caso da evasão no ensino superior do instituto federal do maranhão. *Revista Brasileira de Farmacia*, 10.

Digiampietri, L., Nakano, F., and Lauretto, M. (2016). Mineração de dados para identificação de alunos com alto risco de evasão: Um estudo de caso. *Grad+ Revista de Graduação da USP*, 1:17–23.

Faceli, K., Lorena, A. C., Gama, J., and de Carvalho, A. C. (2011). *Inteligência Artificial: uma Abordagem de Aprendizado de Máquina*. LTC - Livros técnicos e científicos Editora Ltda.

Figueiredo, N. G. d. S. and Salles, D. M. R. (2017). Educação Profissional e evasão escolar em contexto: motivos e reflexões. *Ensaio: Avaliação e Políticas Públicas em Educação*, 25:356 – 392.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Clustering validity checking methods: part ii. *SIGMOD Rec.*, 31 (3):19–27.

Johann, C. C. (2012). Evasão escolar no instituto federal sul-rio-grandense: um estudo de caso no campus passo fundo. Master's thesis, Universidade de Passo Fundo, Passo Fundo-RS.

Lanes, M. and Alcântara, C. (2018). Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1921.

Manhães, L. M. B., Da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., and Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)*, volume 1.

Paz, F. and Cazella, S. (2017). Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 6, page 624.

Plasencia, R. (2011). Desercion universitaria preocupa al mundo. http://www.logrosperu.com/noticias/actualidad/por-desercion-universitaria.html.

Schapire, R. E. (1999). A brief introduction to boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, pages 1401–1406, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tudela, H. V. (2014). Una aproximación teórica a la deserción estudiantil universitaria. *Revista Digital de Investigación en Docencia Universitaria (RIDU)*, 8(1):4.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 4th edition edition.